STUDENT ID NO

MULTIMEDIA UNIVERSITY

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2019/2020

TNL3221 – NATURAL LANGUAGE PROCESSING

(All Sections / Groups)

29 February 2020
2.30p.m. – 4.30p.m.
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of 7 pages with 4 questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) Briefly describe two differences between information retrieval (IR)-based question answering and knowledge-based question answering paradigm. [2 marks]

(b) Given the following contingency tables where "Truth"' is the true class and "Call" is the decision of the classifier. Answer the following questions and show your answer to three decimal places.

Class 1		
	Truth: YES	Truth: NO
Call: YES	180	50
Call: NO	60	200

Class 2		
	Truth: YES	Truth: NO
Call: YES	350	40
Call: NO	30	160

Class 3		
	Truth: YES	Truth: NO
Call: YES	100	20
Call: NO	70	110

i. Compute recall for Class 1.
ii. Compute accuracy for Class 1.
iii. Compute precision for Class 1.
iv. Compute precision for Class 2.
v. Compute precision for Class 3.
vi. Build a pooled contingency table that combines all three classes.
vii. Compute microaverage precision.
viii. Compute macroaverage precision.
ix. What is the potential drawback of microaverage precision?
x. Which precision is more suitable (macroaverage / microaverage) when performance of all the classes is equally important?

[0.5+0.5+0.5+0.5+0.5+0.5+0.5+0.5+0.5+0.5=5 marks]

(c) Austin claimed that the utterance of any sentence in a real speech situation constitutes three kinds of acts. Briefly explain these three kinds of acts. [1.5 marks]

Continued...

(d) Table 1 below shows the Bigrams frequencies of frequencies. Fill in the blanks by calculating the smoothed count c^* based on **Good-Turing estimates**.

[1.5 marks]

c (MLE)	N_c	c^* (GT)
0	74,513,701	
1	37,365	
2	5,820	
3	2,111	
4	1,067	
5	719	
6	468	
7	330	
8	250	
9	179	-

Table 1: Bigram Frequencies of Frequencies

Continued…

QUESTION 2

(a) Write a regular expression for each of the following:

i. That matches the strings: "email" or "e-mail".

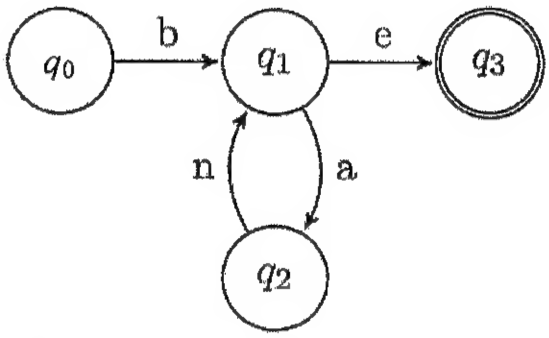
ii. That matches processor speed (examples: MHz, megahertz, Megahertz, GHz, gigahertz, Gigahertz).

iii. That matches strings end with "ed" or "ing".

iv. That matches 11-digit or 12-digit phone numbers with country code and hyphen (examples: +6012-1234567, +6510-12345678)

[1+1+1+1=4 marks]

(b) Based on the following automaton:



i. Write the regular expression.

ii. List the five parameters (Q, Σ, q_0, F, δ) of the automaton.

iii. Draw a state-transition table.

iv. Justify whether it is a deterministic finite state automaton (DFSA) or non-deterministic finite state automaton (NFSA).

[0.5+2.5+2+1=6 marks]

Continued...

QUESTION 3

(a) Given the following context-free grammar:

S → NP VP NP → Nominal NP → Det Nominal Nominal → Noun VP → Verb NP PP PP → Prep NP	Noun → he, she, sea, seashell, seashore Verb → read, saw, learn Det → a, an, the Prep → in, on,for

i. Draw a tree structure for the phrase "she saw seashell on the seashore".
ii. Show the shift-reduce parsing of the phrase "she saw seashell on the seashore".

[2+3=5 marks]

(b) Based on the Levenshtein distance with insertion cost 1, deletion cost 1 and substitution cost 2,

i. Compute the edit distance of "industry" to "interest". Show your work using the edit distance grid.
ii. Compute the edit distance of "dialogue" to "dialect". Show your work using the edit distance grid.

[2.5+2.5=5 marks]

Continued...

QUESTION 4

(a) Given the tag transition probabilities in Table 2, word likelihood probabilities in Table 3 and the part-of-speech tags: Janet/NNP, will/MD, the/DT and bill/NN.

i. List the possible part-of-speech tags for the word "back".

ii. Calculate and justify the best tag for the word "back".

[1+3=4 marks]

	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

Table 2: Tag transition probabilities. Rows are labeled with the conditioning event; thus P(VB|MD) is 0.7968.

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0.000097	0
NN	0	0.000200	0.000223	0.000006	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Table 3: Word likelihood probabilities.

(b) Given the following short product reviews as the training set, each labelled with either positive or negative class:

Review	Text	Class
1	good product fast delivery	+
2	worthy very good quality	+
3	great product	+
4	expensive slow delivery	-
5	bad product lousy quality	-
6	bad quality expensive	-

i. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods. Compute the most likely class for the test review D "good quality very fast delivery"

ii. Justify the class of the test review D.

[3+1=4 marks]

Continued...

(c) Given the following unigram counts (Table 4) and bigram counts (Table 5):

I	Want	To	Eat	Chinese	Food	Lunch	Thai
3437	1215	3256	938	213	1506	459	315

Table 4: Unigram Counts

	I	Want	To	Eat	Chinese	Food	Lunch	Thai
I	8	1087	0	13	0	0	0	0
Want	3	0	786	0	6	8	6	8
To	3	0	10	860	3	0	12	5
Eat	0	0	2	0	19	2	52	28
Chinese	2	0	0	0	0	120	1	0
Food	19	0	17	0	0	0	0	0
Lunch	4	0	0	0	0	1	0	0
Thai	3	0	0	0	0	150	2	0

Table 5: Bigram Counts

i. Compute the bigram probabilities P(Want|I), P(To|Want), P(Eat|To), P(Lunch|Eat), P(Thai|Eat), P(Food|Thai), P(Chinese|Eat), and P(Food|Chinese). Round your answers to two decimal places.

ii. Suppose P(I|<s>) = 0.25, P(</s>|Lunch) = 0.35 and P(</s>|Food) = 0.68, calculate the probability of the sentences "I Want To Eat Thai Food" and "I Want To Eat Chinese Food". Round your final answer to five decimal places.

[1+1=2 marks]

End of Paper